IMPLEMENTACION DE ELEMENTOS DE ANALÍTICA PARA MEJORAR EL PROCESO DE AUDITORIA DE PAGOS TRANSACCIONALES EN LINEA EVALUANDO FACTORES COMO: OPORTUNIDAD Y ALCANCE

OMAR JAVIER SOLER OBANDO

LEIDY JOHANA SANABRIA FARFÁN

FACULTAD DE INGENIERÍA - MAESTRÍA EN ANALÍTICA APLICADA

GONZALO ENRIQUE MEJÍA DELGADILLO – DIRECTOR DE MAESTRÍA

UNIVERSIDAD DE LA SABANA

PÁGINA DE ACEPTACIÓN

_____

Tutor: Rodolfo Yesid Meza Patacón

_____

Jurado:

_____

Jurado:

_____

Jurado:

TABLA DE CONTENIDO

1. LISTA DE TABLAS E ILUSTRACIONES

2. PALABRAS CLAVE

Auditoría de Pagos, Transacciones en Línea, Análisis de Datos, Modelos de Clasificación, Seguridad de Pagos, Eficiencia de Procesos, Análisis Exploratorio de Datos (AED), Power BI.

3. RESUMEN

En la era de la economía digital, la gestión eficiente de los pagos transaccionales en línea es crucial. Esta investigación se centra en la implementación de herramientas analíticas avanzadas y en la adopción de una auditoría basada en riesgos para optimizar la eficiencia y seguridad de estos procesos. Utilizando datos de una empresa líder en servicios de pago en línea, el estudio se enfoca en identificar patrones y anomalías en las transacciones, enfocándose en aquellas aprobadas y declinadas

El análisis exploratorio de datos revela tendencias y factores influyentes, mientras que los modelos de clasificación proporcionan una perspectiva detallada sobre las características de las transacciones. Se desarrolla un visualizador en Power BI, diseñado para resaltar factores críticos, identificando las deficiencias de control y facilitando la toma de decisiones basada en datos. Además, se propone la aplicación de técnicas de estadística multivariada para mitigar transacciones no exitosas y generar alertas tempranas, así como la identificación de los hallazgos de auditoría.

La aplicación de estos métodos analíticos ha permitido identificar varios aspectos clave que influyen en la eficiencia y seguridad de los pagos en línea, tales como el volumen y la frecuencia de las transacciones, así como la clasificación de riesgo

asociada a cada una. Los modelos de clasificación, en particular, han demostrado en su mayoría ser efectivos en la caracterización de transacciones aprobadas y declinadas, proporcionando una base sólida para implementar mejoras proactivas en el proceso de auditoría. Estos resultados preliminares sugieren que la integración de analítica avanzada y auditoría basada en riesgos puede ser un enfoque prometedor para fortalecer la gestión de pagos transaccionales en línea en un entorno económico cada vez más digital.

4. RESUMEN GRAFICO

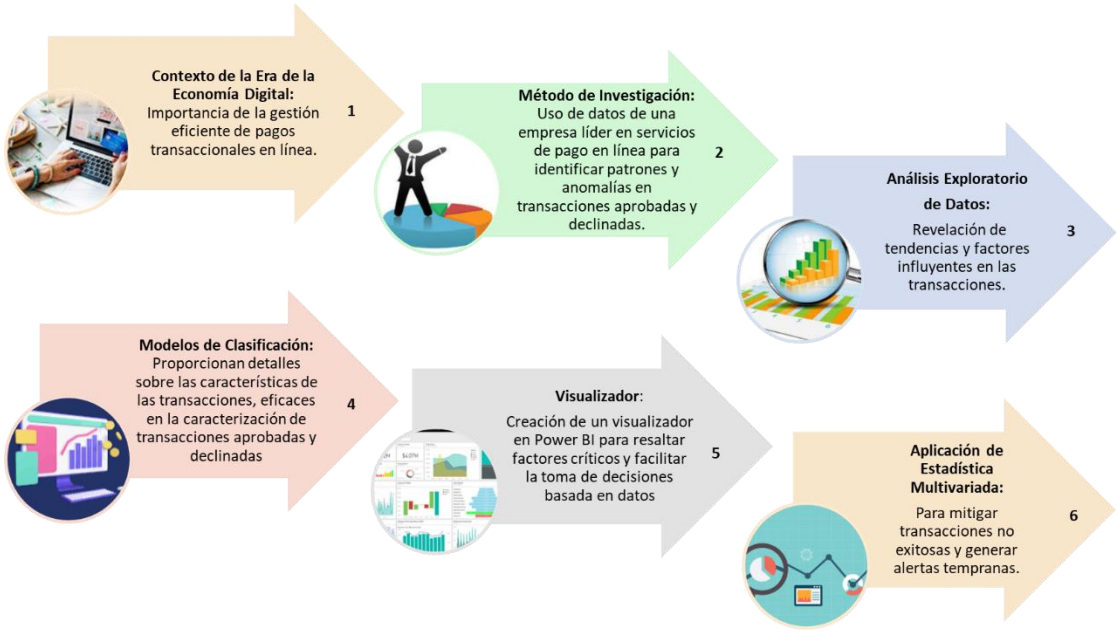

Ilustración 1: Resumen Gráfico

5. INTRODUCCIÓN

En la era actual, caracterizada por la economía digital en auge, la gestión de pagos transaccionales en línea ha emergido como un componente crítico para el éxito de las empresas y la seguridad del consumidor. Este proyecto se enfoca en el ámbito de los pagos en línea abordando el desafío de optimizar la auditoria de este proceso, sector que ha experimentado un crecimiento exponencial y enfrenta desafíos únicos en términos de eficiencia y seguridad, estos desafíos requieren enfoques innovadores para asegurar transacciones fluidas y seguras, adaptándose a un mundo lleno de cambios tecnológicos, transformando las transacciones de efectivo a electrónicas, impulsada por la conectividad de las TIC y la creciente adopción de Fintech, que está redefiniendo el panorama de los negocios financieros y sus operaciones. La pandemia de COVID-19, como se destaca en el Informe Mundial de Pagos 2023 (Capgemini,2023), aceleró esta tendencia, aumentando significativamente el uso de pagos digitales y fomentando niveles más altos de colaboración e innovación en la industria. Esta nueva normalidad ha resaltado la necesidad de una transformación digital rápida en la industria de pagos para mantenerse competitiva, abordando desafíos como la privacidad, la ciberseguridad y el fraude.

En un contexto global, se está observando una transformación significativa en el ámbito de los pagos transaccionales, según la compañía McKinsey se espera que los ingresos por pagos alcancen aproximadamente $2.5 trillones para 2025 (Ilustración 2), con un crecimiento anual del 6% a 7%, un aspecto notable es el descenso global del 16% en los pagos en efectivo en 2020, acompañado de un aumento del 6% en transacciones no monetarias durante el mismo período, este incremento en los pagos en línea resalta la importancia creciente y la necesidad de mejoras en este sector, razón por la cual este estudio se centra en datos de una empresa líder en servicios de pago en línea para identificar patrones y anomalías

en las transacciones, lo que subraya la relevancia de explorar cómo estos cambios están impactando la seguridad y eficiencia de las transacciones financieras (McKinsey & Company ,2021).

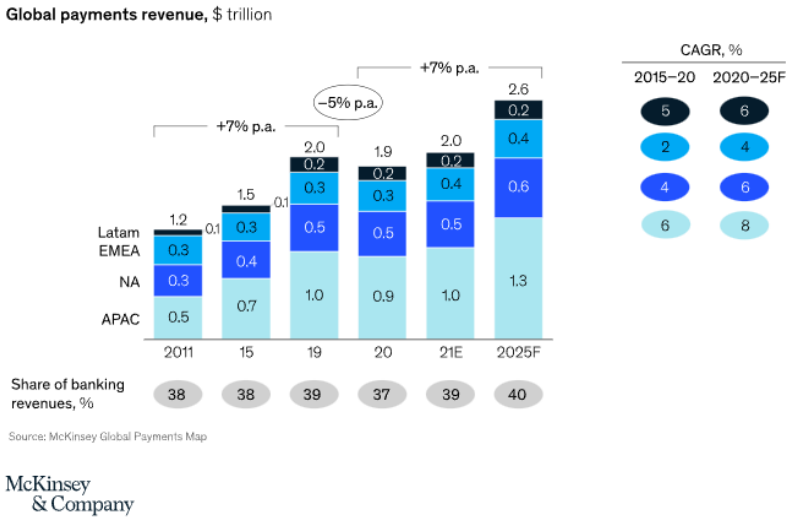

Ilustración 2: Global Payments Revenue. Fuente: McKinsey & Company (2021). Global Payments 2021: Transformation amid turbulent undercurrents. Retrieved from https://www.m ckinsey.com/industries/financial-services/our-insights/global-payments-2021-transformation-amid-turbulent-undercurrents

Uno de los objetivos de este proyecto busca optimizar el proceso de auditoría de pagos transaccionales en línea, enfocándose en la implementación de herramientas analíticas avanzadas y una auditoría basada en riesgos, esta estrategia responde a las necesidades crecientes de eficiencia operativa y seguridad en las transacciones financieras, las cuales han adquirido una importancia crucial en el ámbito empresarial y global. La compañía Deloitte (Una de las llamadas “Big4”: las cuatro firmas de auditoría y consultoría más grandes y prestigiosas del mundo) enfatiza la necesidad de adaptación en la industria de pagos, destacando cómo la pandemia de COVID-19 ha acelerado el cambio hacia los pagos digitales, llevando a las empresas a expandirse a nuevos mercados y enfrentar riesgos más complejos (Deloitte, 2020).

Además, la digitalización de pagos está redefiniendo las interacciones económicas, con un creciente volumen de transacciones que se realizan a través de medios digitales. Esto no solo implica una transformación en la forma en que las empresas y los consumidores realizan transacciones, sino que también plantea nuevos desafíos en términos de ciberseguridad, fraude y cumplimiento normativo. Por lo tanto, es esencial que las auditorías internas en la industria de pagos adopten un enfoque proactivo, utilizando datos y análisis para identificar patrones y anomalías que puedan indicar riesgos potenciales y/o debilidades en el ambiente de control interno.

Este proyecto busca abordar estos desafíos mediante el uso de análisis exploratorio de datos y modelos de clasificación para examinar las transacciones. Se espera que el desarrollo de un visualizador en Power BI mejore la toma de decisiones, proporcionando perspectivas claras y accionables que pueden ayudar a mitigar transacciones fallidas. Esta aproximación no solo busca mejorar la experiencia del cliente, sino que también apunta a fortalecer la infraestructura y la operatividad de los sistemas de pago en línea, testeando su seguridad y eficacia en un mundo cada vez más digitalizado.

En resumen, este proyecto no solo busca abordar un problema crucial en el contexto empresarial moderno, sino que también apunta a generar un impacto significativo a nivel de la compañía, fortaleciendo las prácticas de auditoría y la gestión de pagos transaccionales en un mundo cada vez más digitalizado.

6. PREGUNTA DE INVESTIGACIÓN APLICADA

¿Cómo puede la implementación de elementos de analítica y la metodología de auditoría basada en riesgos contribuir a mejorar la eficiencia del proceso de auditoría de pagos transaccionales en línea?

7. MARCO CONCEPTUAL

El marco conceptual de este proyecto proporciona la base teórica y el contexto necesario para comprender y abordar los desafíos relacionados con la auditoría de pagos transaccionales en línea, este marco conceptual comprende dos partes: un marco teórico que examina conceptos claves y un estado del arte que revisa de manera exhaustiva las prácticas y teorías existentes.

La Auditoría Basada en Riesgos (ABR), es una metodología que se ha cobrado gran importancia en el contexto empresarial actual. Según (Pickett, 2015), la ABR no solo se centra en identificar y evaluar áreas de alto riesgo, sino que también es clave para asegurar una gobernanza efectiva, la eficiencia de controles y la gestión de riesgos en las organizaciones.

Las fases clave de una auditoría según las Normas Internacionales, son la planificación, la ejecución y el reporte**.** La planificación es la etapa más importante, ya que aquí se definen los riesgos a evaluar, los procedimientos a seguir y el enfoque de auditoría. Un plan de auditoría bien estructurado permite anticipar riesgos y asignar los recursos de manera eficiente. En este proyecto, la fase de planificación mejora significativamente con la integración de herramientas de análisis de datos, permitiendo una selección más precisa de las áreas de riesgo y procedimientos de auditoría basados en datos concretos. La ejecución, que abarca la aplicación de los procedimientos definidos en la planificación, se refuerza mediante el análisis exploratorio de datos, facilitando la identificación de anomalías y patrones en grandes volúmenes de transacciones. Finalmente, el reporte se mejora al presentar resultados más claros, respaldados por evidencia analítica, mejorando así las recomendaciones a implementar en el control interno.

Esta metodología, como se detalla, permite una asignación eficiente de recursos de auditoría, priorizando áreas críticas para garantizar la integridad y fiabilidad de los

sistemas y transacciones. Permitiendo que, la planificación, ejecución y reporte del proceso de auditoría se vean enriquecidos y mejorados gracias a la incorporación de técnicas de analítica avanzada.

Además, es esencial considerar la sensibilidad de los auditores al riesgo estratégico en el contexto de la auditoría basada en riesgos, pues un estudio relevante en este campo sugiere que los auditores pueden subestimar los riesgos estratégicos en información de bajo riesgo. Asimismo, el estudio demuestra que anticipar las estrategias de los gerentes puede ser clave para una auditoría eficaz, lo que es una perspectiva fundamental para nuestro trabajo. (Bowlin, 2011).

El enfoque de auditoría continua en entornos de big data, destacado en el trabajo de Kiesow, Zarvic y Thomas, es particularmente pertinente para este análisis, pues esta metodología, que incorpora Herramientas de Pruebas de Auditoría Asistidas por Computadora (CAATTs), permite a los auditores manejar y analizar extensos conjuntos de datos de transacciones de manera eficiente y continua. La auditoría continua se adapta a la dinámica cambiante y a menudo compleja del entorno de pagos en línea, ofreciendo una supervisión más profunda y oportuna de las transacciones. La integración de CAATTs facilita la identificación rápida de patrones inusuales, lo que mejora significativamente la detección de riesgos y la toma de decisiones en tiempo real, la aplicación de este enfoque puede aportar en el análisis del proceso de pagos, permitiendo responder dinámicamente a los desafíos emergentes en un entorno de rápida evolución (Kiesow et al., 2014).

El proceso de pagos transaccionales en línea es una secuencia compleja y multifacética que involucra a varias partes interesadas, este proceso comienza cuando el cliente seleccionando un producto o servicio y proporciona los detalles de pago, generalmente a través de una tarjeta. La transacción es transmitida a la franquicia de la tarjeta, luego al banco emisor para su validación, y finalmente regresa al comercio con una respuesta de aprobación o rechazo. Aunque para el

cliente parece un proceso instantáneo, cada paso está sujeto a múltiples verificaciones para garantizar tanto la seguridad como la eficiencia, lo que resulta fundamental para su análisis en el contexto de la auditoría basada en riesgos. (Rambure, D., & Nacamuli, A, 2008).

La confidencialidad de los datos es un aspecto esencial en este contexto, la naturaleza sensible de los datos financieros requiere un manejo cuidadoso y medidas de seguridad robustas, por lo que la gestión adecuada de la privacidad y la seguridad de los datos es crucial para mantener la confianza del cliente y cumplir con las regulaciones. En este proyecto, se prestará especial atención a la confidencialidad de los datos proporcionados por la empresa de servicios de pagos en línea, garantizando que toda la analítica se realice respetando los estándares de privacidad y seguridad. (Han et al., 2012)

Al integrar la auditoría basada en riesgos con el proceso de pagos transaccionales en línea, se crea un enfoque holístico que fortalece la integridad de las transacciones. Utilizando técnicas avanzadas de análisis de datos en combinación con la ABR, es posible identificar riesgos potenciales en tiempo real, esto permite a las organizaciones responder de manera proactiva, ajustando los controles y medidas de seguridad, para asegurar que tanto el cliente como el comercio disfruten de una experiencia de pago segura y confiable.

A continuación, mostramos el proceso actual de la compañía sobre el que la auditoria es crucial:

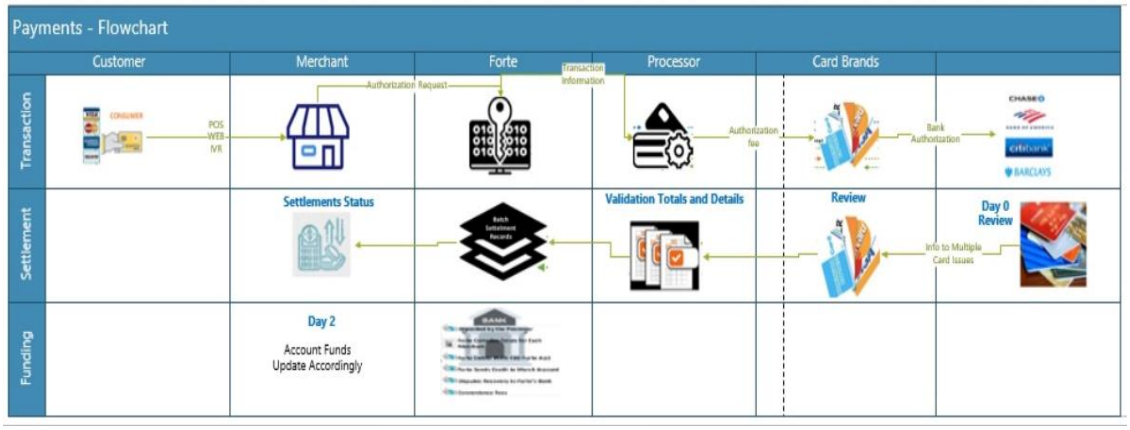

Ilustración 3:Conceptualización de la Auditoría de Pagos Transaccionales en Línea del caso en estudio

La ciencia de datos juega un papel esencial en la revisión del proceso de pagos, ya que proporciona herramientas avanzadas para el análisis y la toma de decisiones estratégicas. A través de técnicas de minería de datos, se pueden analizar grandes conjuntos de datos, lo cual es crucial para realizar un análisis profundo, mejorando la eficiencia y manteniendo la seguridad del proceso. Este enfoque basado en la aplicación de pensamiento analítico robusto y en la solución de problemas dentro del contexto empresarial, permite realizar una gestión de riesgos y una auditoría más efectiva, orientada a los desafíos actuales del sector (Provost & Fawcett, 2013).

Revisando el concepto de minería de datos se identifica que es un componente clave de la analítica, ya que implica el proceso de descubrir patrones, tendencias y extraer información valiosa de conjuntos de datos. En el ámbito de los pagos en línea, este enfoque permite identificar relaciones y anomalías en las transacciones, ofreciendo puntos críticos para mejorar la seguridad y la eficiencia de los procesos de pago, de igual forma, es importante incluir métodos avanzadas de clasificación, agrupación y detección de anomalías, ya que estos son especialmente útiles para

identificar comportamientos inusuales, así como para comprender mejor las tendencias del cliente. (Han et al., 2012),

Es indispensable iniciar la revisión de los datos realizando un Análisis Exploratorio de Datos (AED), ya que en este contexto es crucial para mostrar en detalle el comportamiento de las variables, lo que permite una comprensión más profunda y precisa de las transacciones de pagos en línea. Al utilizar el AED, podemos analizar estructuras complejas de datos, identificando tendencias no evidentes y áreas potenciales de riesgo. Este enfoque es particularmente relevante, pues permite conocer como varían las transacciones, y como están sujetas a una gama de factores. Por ello, es necesario asegurar que el análisis sea exhaustivo y este orientado a descubrir aspectos relevantes que puedan influir el proceso de pagos (Liu, 2014).

La clasificación es una técnica del aprendizaje automático que asigna etiquetas a nuevos datos, basándose en características clave que permiten agruparlos en categorías. Su objetivo es organizar grandes volúmenes de datos para extraer conocimiento relevante. Existen diferentes tipos de algoritmos de clasificación, cada uno con sus propias fortalezas y debilidades, dependiendo del tipo y la complejidad de los datos que se estén analizando. Los algoritmos más comunes incluyen árboles de decisión, máquinas de soporte vectorial y redes neuronales (Zaki, 2020).

Para medir qué tan bien funcionan los modelos de clasificación, se utilizan varias métricas importantes. La exactitud (Acurracy) indica el porcentaje de predicciones correctas que hace el modelo en general. La precisión nos dice cuántas de las predicciones positivas fueron realmente correctas, mientras que la sensibilidad (o recall) muestra cuántos de los casos positivos reales fueron identificados correctamente por el modelo. Por último, la puntuación F1 es una combinación de precisión y sensibilidad, que ayuda a equilibrar ambas métricas, especialmente

cuando hay un desbalance entre las clases. Estas métricas son esenciales para entender cómo mejorar los modelos de clasificación (Erickson & Kitamura, 2021).

Junto con las métricas es importante revisar algunos algoritmos de clasificación y técnicas de preprocesamiento:

Regresión logística es un modelo estadístico que se utiliza para tareas de clasificación binaria. Estima la probabilidad de que una instancia pertenezca a una de las dos clases. (Pedregosa et al., 2011, Sección Regresión logística).

Los Árboles de Decisión son algoritmos de clasificación y regresión que utilizan una estructura de árbol para tomar decisiones basadas en características. Son altamente interpretables y manejan tanto datos categóricos como numéricos. En grandes volúmenes de datos, los árboles de decisión pueden adaptarse rápidamente y proporcionar resultados precisos, siendo una herramienta valiosa para análisis exploratorios y modelos predictivos robustos (Loh, 2014).

Como ventajas, se menciona la simplicidad e interpretabilidad: Los árboles de decisión son fáciles de entender y visualizar y tienen la capacidad para manejar datos categóricos y numéricos, sin embargo, frecuentemente presentan propensión al sobreajuste, es decir los árboles de decisión pueden sobre ajustar los datos de entrenamiento si no se podan adecuadamente.

Los bosques aleatorios (Random Forest) modelo que funciona construyendo una multitud de árboles de decisión durante el tiempo de entrenamiento, para tareas de clasificación, la salida del bosque aleatorio es la clase seleccionada por la mayoría de los árboles, es efectivo en una amplia variedad de tareas y pueden manejar datos tanto categóricos como numéricos. Los bosques aleatorios también son relativamente robustos a los valores atípicos (Breiman, 2001).

XGBoost, algoritmo de aprendizaje automático de refuerzo por árboles de decisión escalable, el cual se basa en la idea de construir secuencialmente árboles de decisión, donde cada árbol se entrena para corregir los errores de los árboles

anteriores, tiene la capacidad de manejar conjuntos de datos grandes y con múltiples características. (Chen & Guestrin, 2016).

Smote: es una técnica de preprocesamiento utilizada para abordar el desequilibrio de clases en un conjunto de datos, la cual crea nuevas instancias de la clase minoritaria a través de combinaciones de instancias vecinas. Funciona seleccionando ejemplos que están cerca en el espacio de características, trazando una línea entre los ejemplos de este y dibujando una nueva muestra en un punto a lo largo de esa línea (Chawla et al., 2002; Chen & Guestrin, 2016). Esto puede ayudar a mejorar el rendimiento de los modelos de clasificación en conjuntos de datos desequilibrados al evitar que el modelo se sesgue hacia la clase mayoritaria.

Por otra parte, Random Undersampling es una técnica de preprocesamiento utilizada para abordar el desequilibrio de clases en un conjunto de datos. Esta técnica ayuda a balancear las clases, mejorando el rendimiento de los modelos de clasificación y evitando que el modelo se sesgue hacia la clase mayoritaria, siendo una técnica simple y rápida de implementar. En ocasiones, la reducción del tamaño de la clase mayoritaria puede llevar a la pérdida de información valiosa con esta técnica. Haixiang et al. (2017 (Haixiang, 2017)).

Abordando otro tema importante de estudio, se evaluará el modelo de aprendizaje automático Isolation Forest, el cual se usa para detectar datos que no siguen un patrón normal, es decir, "anomalías". Este modelo es especialmente útil porque puede adaptarse a cambios en los datos conforme estos se van actualizando, lo que lo convierte en una herramienta adecuada para analizar información en tiempo real, donde los datos cambian constantemente (Ding & Fei, 2013).

8. OBJETIVOS

Objetivo General

Optimizar el proceso de auditoría de pagos transaccionales en línea mediante la implementación de elementos de analítica, evaluando específicamente los factores de oportunidad y alcance en el proceso.

Objetivos Específicos

- Realizar un Análisis Exploratorio de Datos para identificar patrones, tendencias y factores que inciden en las transacciones aprobadas y declinadas.
- Utilizar modelos de clasificación para caracterizar y comprender las transacciones aprobadas y declinadas, basándose en los factores identificados a través del AED, y evaluar la precisión de estos modelos.
- Identificar estrategias para mitigar el número de transacciones que no debieron ser declinadas utilizando técnicas de estadística multivariada, proponiendo mejoras y generación de alertas tempranas.
- Desarrollar un visualizador en Power BI que destaque factores críticos en las transacciones, facilitando la comprensión y la toma de decisiones.
- Evaluar la efectividad del sistema de control interno del proceso de pagos transaccionales en línea ejecutando pruebas de auditoría de validación masiva de información por medio de elementos de analítica y compararlo frente a los posibles resultados de auditoría tradicionales.

9. METODOLOGIA

Este proyecto analizó información de pagos transaccionales proporcionada por una empresa líder en servicios de pago en línea. La compañía, cuyo slogan es "Customer Experience, Billing and Payments Solution", se especializa en ofrecer servicios y soluciones empresariales a la medida en la gestión de ingresos, experiencia al cliente y soluciones de pago en industrias como telecomunicaciones, retail, servicios financieros y gubernamentales.

Su enfoque todo en uno de la Empresa permite aceptar cualquier tipo de pago en cualquier dispositivo o canal, lo que ha fidelizado a algunos de sus clientes durante más de 15 años. Sus soluciones de pago están diseñadas con los protocolos de seguridad más avanzados, protegiendo los datos de pago confidenciales y defendiendo a las empresas contra el fraude.

La compañía presta servicio a organizaciones de cualquier tamaño y en cualquier sector, ofreciendo distintos paquetes de precios para empresas pequeñas, medianas y grandes que buscan aceptar pagos y procesar cualquier volumen.

La estructura analítica de la compañía está diseñada para fomentar una cultura de empoderamiento de datos en toda la organización. Cada departamento puede desarrollar sus propios informes, análisis y métricas utilizando los datos disponibles, promoviendo así un enfoque autosuficiente. Este enfoque simplifica el acceso a los datos, mejora la agilidad y la toma de decisiones, y optimiza los procesos, resultando en una mayor eficiencia.

9.1 Entendimiento de la base de datos

Las bases de datos proporcionadas provienen de la tabla “Daily Transaction”, un archivo que contiene datos de todos los pagos diarios procesados a través del sistema transaccional de la compañía en todos los países donde opera. Las transacciones están registradas en horario "Central Time". Este archivo contiene información detallada sobre transacciones realizadas durante los años 2022 y 2023,

con un total aproximado de 2.2 millones de transacciones acumuladas por trimestre (Ilustración 4).

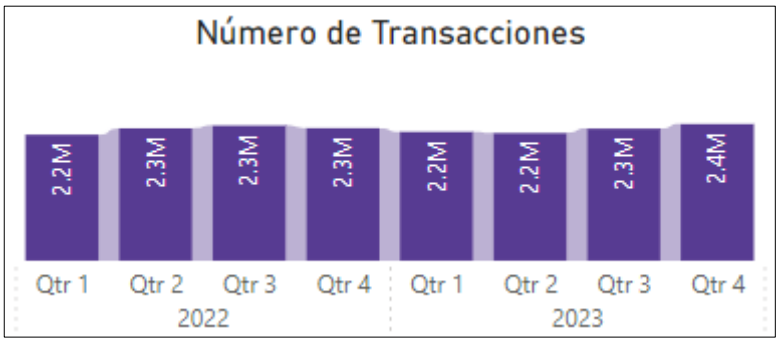

Ilustración 4: Total de transacciones trimestrales procesadas en el sistema transaccional (Cifras expresadas en millones).

Para este análisis, se seleccionaron específicamente los meses de enero, junio, agosto y diciembre de ambos años (Ilustración 5). Estos meses fueron elegidos debido a su relevancia particular en términos de volumen y valor de transacciones, así como por su importancia en los ciclos financieros y comerciales globales. Enero y junio son considerados meses clave debido al inicio del año fiscal y al cierre del primer semestre, períodos en los que se realizan pagos de impuestos y ajustes contables en varios países (Investopedia, s.f.; Finance, s.f.). Agosto destaca por ser un mes de reactivación comercial tras las vacaciones de descanso escolar y laboral. Diciembre por su parte, es conocido por su alta actividad transaccional debido a las compras navideñas y los cierres anuales (Swing, s.f.).

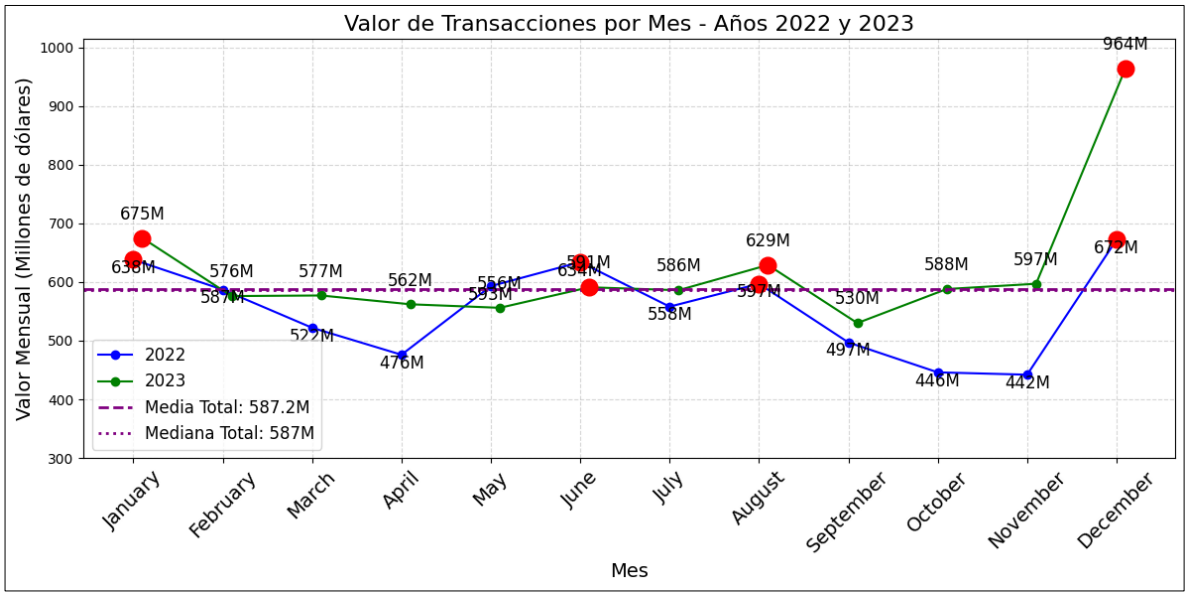

Ilustración 5: Valor mensual de pagos ingresados al sistema transaccional a nivel global (Cifras expresadas en millones).

La ilustración 5 muestra el valor mensual de las transacciones y cierto incremento en los meses seleccionados convirtiéndolos en períodos claves para el análisis. Esto permite capturar una visión comprensiva del comportamiento transaccional durante los meses más activos del año, proporcionando perspectivas valiosas para la auditoría y los análisis a ejecutar.

Los datos correspondientes a los 8 meses seleccionados se recibieron en archivos independientes de formato .CSV no estructurados, uno por cada día del mes. En total se consolidaron 246 archivos que contienen 6,503,910 transacciones. Para alinear el análisis de los datos transaccionales con la auditoría basada en riesgos y el cumplimiento de los pagos, se realizó un cruce con una tabla paramétrica de códigos de categoría de comerciantes (MCC) y el nivel de riesgo definido por la compañía. Este proceso permitió obtener y añadir las variables category_risk y risk_business_description, que aportan información detallada sobre el nivel de riesgo asociado a cada categoría de transacción y la descripción del riesgo del negocio correspondiente. A continuación, se presenta en la tabla 1 la descripción de cada campo, su tipo de dato, ejemplos representativos y estadísticas relevantes,

como la cantidad de valores únicos y registros nulos, para proporcionar una visión más clara del contenido de los datos.

Campo	Descripción del Campo	Tipo	Ejemplo	Valores Únicos	Registros Nulos
Amount Total	Valor total de la transacción.	Numérico, en dólares	Mínimo --815.76 y Máximo 10000000.	n/a	0
Auth Results Detail	Código y descripción del estado de la transacción.	Categórico	[A01:Approved;A01:Test Approval;U10:Duplicate Transaction U20:Invalid Credit Card Number;....]	202	0
Auth Results Type	Identifica si la transacción es aprobada (A) o declinada (D)	Categórico	[A, D]	2	0
Card Category	Categoría de la tarjeta de crédito utilizada para la transacción	Categórico	[Classic, Platinum, Gold,…..]	207	571,202
Card Type	Tipo de tarjeta utilizada para la transacción	Categórico	[DEBIT CREDIT CHARGE CARD DEBIT OR CREDIT nan]	4	0
Category Risk	Categoría de riesgo del negocio.	Categórico	[Low Moderate High Prohibited Restricted]	5	5,544
Cvv Result	Código de seguridad de 3 o 4 dígitos de las tarjetas.	Categórico	Campo confidencial	7	1,907,507
Entry Type	Franquicia de la tarjeta.	Categórico	[Visa, Mast, Amer, Loca Rupa……]	17	0
Issuing Country	País en el que se realizó la transacción	Categórico	[CAN BRA BRB BRN CAN CHE CHL CHN ESP FRA GBR GEO GRC GRD GUM USA.....]	184	0
Location Id	ID de la ubicación del comerciante que proceso la transacción.	Categórico	[107625 234041 186138 ... 332172 333066 334577]	48,763	0
Mcc	Actividad Económica del comercio,	Categórico	[4119, 8099, 7399. 7393, 8398, 8299…..]	202	7,059
Qty	Cantidad de transacciones realizadas.	Numérico	Mínimo 1 y Máximo 76452	n/a	0
Received Date	Fecha en la que se realizó la transacción.	Fecha	[2022-01-01; 2022-01-02; 2022-01-03; 2023-01-03;….]	246	0
Risk Business Description	Descripción del negocio	Categórico	[Ambulance Services Equipment, Software Book Stores…..]	187	5,544
Transaction Type	Tipo de transacción	Categórico	[10 11 13 15]	4	0
Mes	Mes y año en el que se realizó la transacción.	Fecha	[2022-01;2023-01;2022-06;2023-06;2022-08;2023-08;2022-12; 2023-12]	8	0

Tabla 1: Descripción de la Base Analizada

9.2 Análisis Exploratorio

Dando continuidad al proceso de consolidación y validación de los datos, se procedió con el análisis exploratorio para identificar posibles inconsistencias, valores faltantes y patrones relevantes que pudieran afectar la calidad del análisis. A continuación, se presenta un resumen de los hallazgos más importantes y las decisiones tomadas para depurar y preparar los datos antes de aplicar los modelos analíticos:

Durante el análisis de valores faltantes, se identificaron 7,059 registros nulos en el campo MCC de un total de 6,503,910 registros en la base de datos, debido a que 43 location_id no tienen actividad económica asignada. Además, se encontraron 5,544 registros nulos en category_risk, correspondientes a 6 actividades económicas sin categoría de riesgo asignada. Estos resultados sugieren un posible hallazgo de auditoría, ya que los valores nulos podrían deberse a la falta de actualización de las bases de datos con los cambios de los listados de MCC y category_risk. Por lo anterior, y considerando que estos valores nulos representan un posible error de actualización, se procedió a retirar dichos registros para garantizar la calidad del análisis y asegurar la correcta aplicación de los modelos analíticos.

De igual forma, se identificó que las columnas card_category y cvv_result, presentaban 8.78% y 29.25% de valores nulos, respectivamente. Estas columnas no impactan significativamente los resultados de las transacciones, ya que la mayoría de las decisiones de autorización o denegación se toman con éxito independientemente de la categoría de la tarjeta. En el caso de cvv_result, este valor no se solicita en todos los canales, como en transacciones presenciales, o compras recurrentes, y además corresponde a un dato confidencial. Sin embargo, mantener estas columnas supone un desafío en términos de calidad y completitud de los datos, lo que limita la capacidad de obtener conclusiones claras y precisas. Por esta razón, se excluyeron del análisis para enfocar el trabajo en variables con mayor integridad de datos. Los registros eliminados se conservaron para realizar pruebas analíticas de auditoría toda vez que puede obedecer a hallazgos de auditoría.

Durante la exploración de los datos y la validación de montos totales por tipo de transacción, se identificaron dos transacciones declinadas atípicas el 20 de enero de 2022 y el 29 de agosto de 2022, por montos de 6,786 millones y 1,100 millones de dólares respectivamente. Estas transacciones fueron detectadas en el campo de

transacciones declinadas (D), dado a que como se muestra en la tabla 2 estos meses presentaban valores fuera del promedio.

Tabla Comparativa del monto total (en miles) por Mes y tipo de Resultado:

auth_results_type	A	D
mes		
2022-01	601234.00	6848943.00
2022-06	545102.00	87379.00
2022-08	579224.00	1140542.00
2022-12	637260.00	34200.00
2023-01	656220.00	18297.00
2023-06	572202.00	18053.00
2023-08	615029.00	13318.00
2023-12	923669.00	39731.00

Tabla 2: Comparación del monto total por mes y tipo de resultado

Tras la validación con la compañía, sé confirmó que estas transacciones correspondían al agrupamiento de varias operaciones en un mismo día. Por lo tanto, se decidió eliminar estos registros para no alterar el análisis ni los modelos que se implementaran posteriormente.

Después del tratamiento de los datos faltantes y atípicos, la base quedó con un total de 6,491,305 registros y el valor máximo se redujo a 10 millones de dólares luego de estar en 6,786 millones de dólares. Este tratamiento de los datos faltantes y atípicos fue necesario para poder aplicar los modelos analíticos del proyecto. No obstante, los registros eliminados se conservaron para su posterior análisis en las pruebas de auditoría, ya que podrían constituir hallazgos de auditoría.

9.2.1 Análisis Univariado

El análisis univariado permitió explorar la distribución de las variables individuales en el conjunto de datos. Este análisis es crucial para identificar patrones, tendencias y posibles sesgos antes de avanzar hacia los modelos de clasificación. A continuación, se presentan los hallazgos más relevantes de las variables categóricas (location_id, mcc, transaction_type, entry_type, card_type, issuing_country, Category_risk) (Ilustración 6 y 7):

- Location_id: La distribución de location_id muestra una dispersión uniforme con ciertos picos, lo que sugiere que algunos comercios procesan un mayor número de transacciones. Esto puede deberse a factores como el tamaño o la actividad económica del comercio.
- MCC (Merchant Category Code): Se observó una alta concentración de transacciones en ciertos códigos MCC, como el 6513 (Property Management) y el 9399 (Government Services), lo que indica que el sector de gestión de propiedades y los servicios gubernamentales generan un volumen significativo de transacciones. Esto sugiere una concentración de actividad en sectores clave para la compañía.
- Transaction_type: En el tipo de transacción predomina el código 10 (Transacción de Venta), que representa la mayoría de las transacciones, seguido de los códigos 11 (Autorización/Captura) y 13 (Reembolso/Distribución). Esto destaca una fuerte concentración en transacciones relacionadas con ventas directas.
- Entry_type: Las franquicias más representativas en las transacciones son Visa y MasterCard, seguidas por American Express, lo que refleja una alta aceptación y uso de estas marcas por parte de los clientes.
- Card_type: La mayoría de las transacciones se realiza con tarjetas de crédito, aunque las transacciones con tarjetas de débito también tienen una frecuencia significativa. Esto sugiere una preferencia de los clientes por las modalidades de pago a crédito.
- Issuing_country: El análisis de esta variable muestra que, aunque los pagos se pueden realizar globalmente, la mayor parte de las transacciones proviene de Estados Unidos (USA), seguido de Canadá (CAN) y Brasil (BRA). Esto refleja una concentración de actividad transaccional en estas regiones.
- Category_risk: La mayoría de las transacciones se clasificaron en la categoría de bajo riesgo, aunque, una porción considerable también se encontró en la categoría de riesgo moderado, mientras que las categorías de

alto riesgo, prohibido y restringido presentaron una cantidad mínima de transacciones.

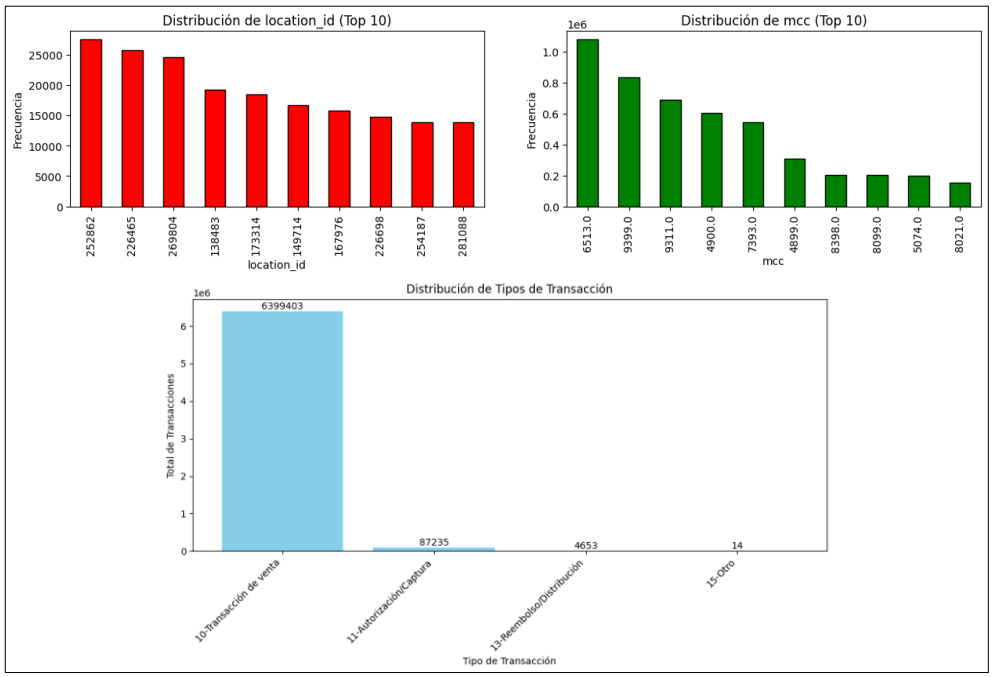

Ilustración 6:Distribución de las Principales Variables Categóricas: Location ID, MCC y Tipos de Transacción

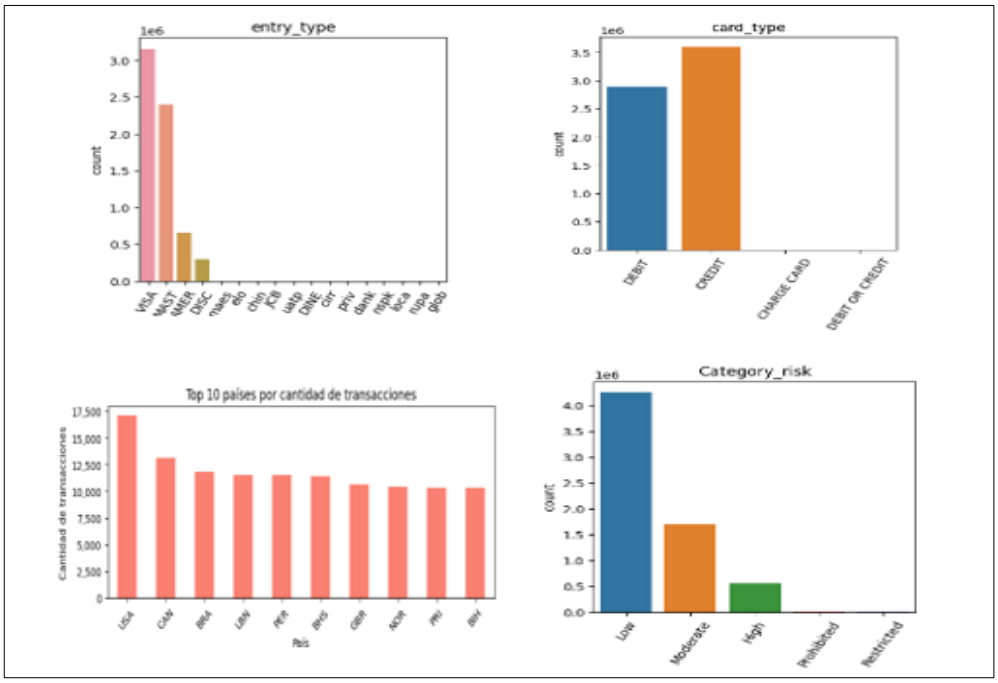

Ilustración 7:Distribución de Variables Categóricas: Franquicia, Tipo de Tarjeta, Países Principales y Categoría de Riesgo

Estos hallazgos proporcionan una comprensión clara de las áreas clave para la gestión del negocio, destacando tres aspectos fundamentales: las preferencias de los clientes (Visa y tarjetas de crédito), la concentración geográfica (con la mayoría de las transacciones en USA, CAN y BRA) y el nivel de riesgo asumido (predominio en transacciones de bajo riesgo). Estos patrones ofrecen insumos relevantes para orientar la gestión de riesgos y facilitar la toma de decisiones informadas en el negocio.

Como parte del análisis univariado, se presentan las estadísticas descriptivas de las variables numéricas amount_total (valor total de la transacción) y qty (cantidad de transacciones) resumidas en la tabla 3:

- Amount_total: El valor total de las transacciones mostró una media de $838.72 USD, con una desviación estándar significativamente alta

($11,451.32 USD), lo que indica una gran variabilidad en los montos de las transacciones. El valor máximo registrado fue de $10,000,000 USD.

- Qty: La cantidad de transacciones mostró una media de 4.38, con una desviación estándar de 71.96, lo que indica que, aunque la mayoría de las transacciones involucraban pocas transacciones, hay algunas transacciones que sobresalen por su volumen. El valor máximo fue de 76,452 transacciones.

Métrica	Amount_total	Qty
Cantidad	6,491,305	6,491,305
Media	838.72	4.38
Desviación estándar	11,451.32	71.96
Mínimo	-815.76	1.00
25% Percentil	52.00	1.00
Mediana (50%)	157.73	1.00
75% Percentil	516.18	2.00
Máximo	10,000,000	76,452

Tabla 3: Estadísticas Descriptivas de las Variables Numéricas: Amount_total y Qty

Además, en la Ilustración 8 se presentan gráficos de boxplots que muestran la distribución de las variables amount_total y qty. Se aplicó transformación logarítmica para reducir la asimetría de los datos y facilitar su interpretación.

La variable amount_total muestra una concentración de transacciones con montos relativamente bajos, con una media cercana a 6 en la escala logarítmica (equivalente a aproximadamente $403 USD en la escala original). Sin embargo, se observan valores altos hacia la derecha, lo que indica la existencia de transacciones con montos altos que pueden tener un impacto considerable en los resultados.

Por otro lado, la variable qty refleja un comportamiento similar, donde la mayoría de las transacciones involucran cantidades pequeñas de productos, pero se identifican

outliers que representan transacciones con cantidades muy altas. Estos casos inusuales podrían corresponder a clientes o comercios específicos que generan un volumen significativo de transacciones en un solo día, y su análisis más profundo podría proporcionar información adicional sobre patrones de compra masiva o eventos especiales.

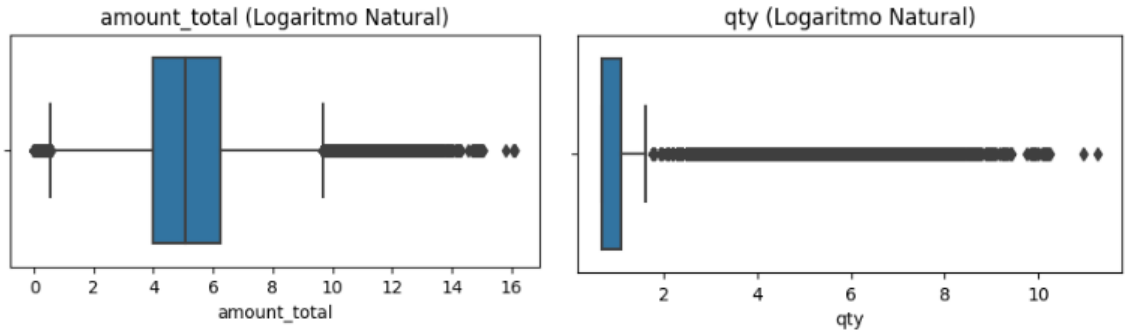

Ilustración 8:Boxplots de las Variables Numéricas

9.2.2 Análisis Multivariado

La relación entre el estado de la transacción (Auth Results Type), las cantidades totales de transacciones aprobadas y declinadas, así como sus respectivos montos se presenta en la Ilustración 9 y 10. A continuación, se destacan los hallazgos más relevantes:

- Diciembre de 2023 presenta el valor más alto de transacciones aprobadas, alcanzando 923,7 millones de USD, mientras que junio de 2022 y enero de 2022 destacan por tener los montos más altos de transacciones declinadas, con 87,4 millones de USD y 62,9 millones de USD, respectivamente.
- El 95,6% de la cantidad de transacciones realizadas fueron aprobadas, representando 27,22 millones de transacciones, mientras el 4,4% restante, equivalente a 1,24 millones de transacciones fueron declinadas.
- El 94,2% del monto total de las transacciones, equivalente a 5.129,94 millones de USD, corresponde a transacciones aprobadas, mientras que el

5.8% restante, equivalente a 314.44 millones de USD, corresponde a transacciones declinadas.

Este análisis muestra un desbalance entre transacciones aprobadas y declinadas. Aunque las transacciones declinadas son menores en cantidad, concentrar montos significativos, por lo que es importante realizar un análisis detallado de este tipo de transacciones, ya que pueden tener impacto en la compañía.

Este análisis destaca la importancia de monitorear continuamente tanto la cantidad como el valor de las transacciones declinadas para asegurar que los procesos de pago funcionen acordes a lo establecido por la compañía. De igual forma, una gestión efectiva del riesgo permitirá mejorar la experiencia del cliente y ayudará a disminuir el riesgo de pérdidas económicas.

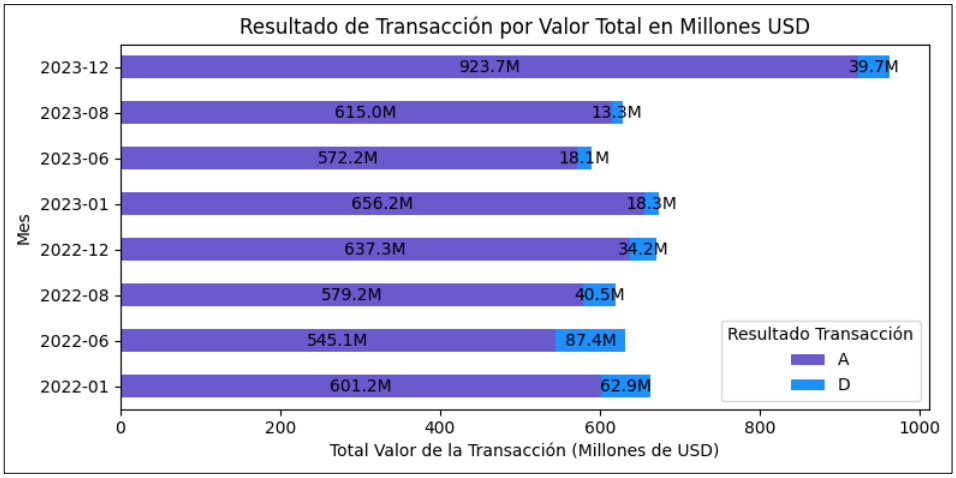

Ilustración 9:Participación de transacciones aprobadas y declinadas por valor mensual.

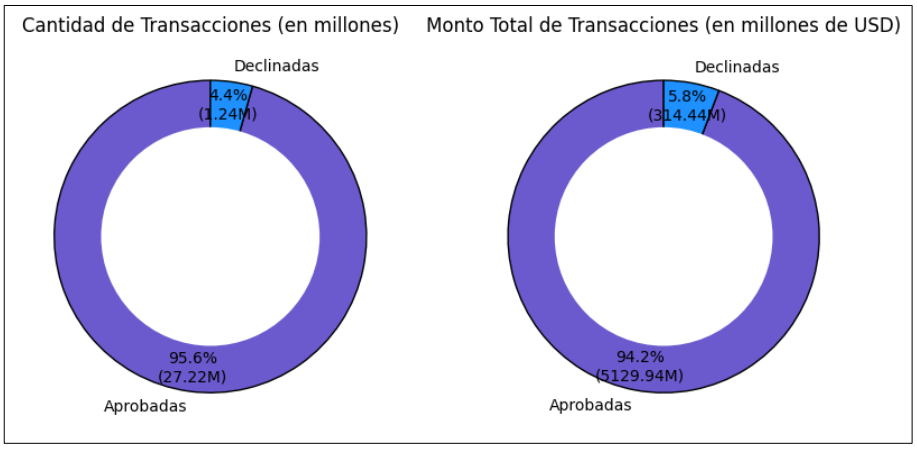

Ilustración 10:Participación de transacciones aprobadas y declinadas por cantidad y monto total

Continuando con el análisis del estado de la transacción, la Ilustración 11 presenta la relación entre el estado de la transacción y las categorías de riesgo En el eje Y se muestra la cantidad total de registros por cada categoría de riesgo: bajo (`low`), moderado (`moderate`), alto (`high`), prohibido (`prohibited`) y restringido (`restricted`), mostrando los siguientes resultados:

- La mayoría de las transacciones, tanto aprobadas como declinadas, se concentran en las categorías de riesgo bajo (`low`) y moderado (`moderate`), lo que refleja una tendencia conservadora en la gestión del riesgo.
- Las categorías de riesgo alto (`high`), prohibidas (`prohibited`) y restringidas (`restricted`), presentan una cantidad mínima de transacciones, sugiriendo que la exposición de la compañía a estos niveles de riesgo es limitada. Las transacciones aprobadas tienen una distribución de riesgo similar a las transacciones declinadas, ya que lambas se concentran en los niveles de riesgo bajo y moderado.

Estos resultados sugieren que el sistema de aprobación de transacciones está funcionando adecuadamente para mitigar el riesgo, manteniendo la mayor parte de las operaciones dentro de categorías de riesgo bajo y moderado, tanto para transacciones aprobadas como declinadas.

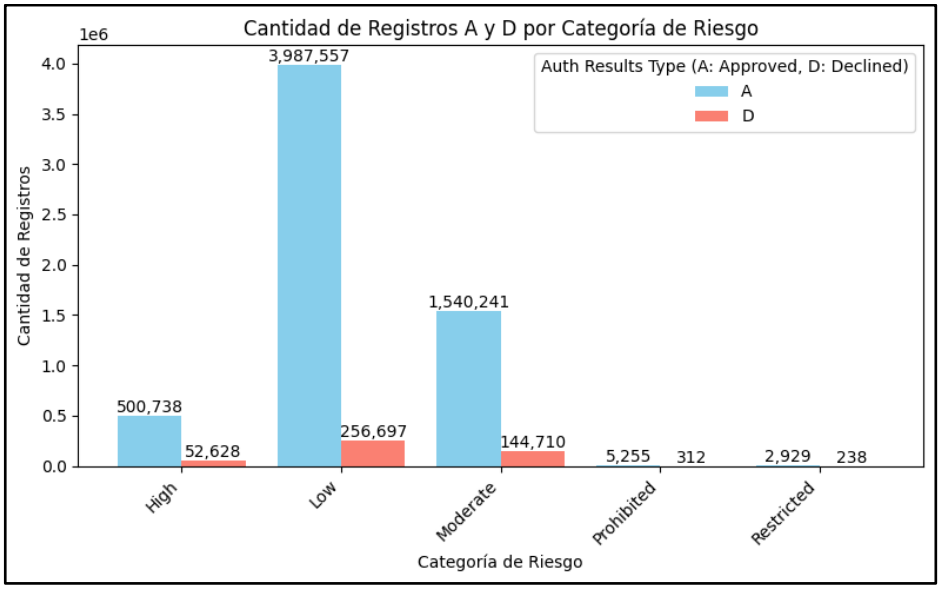

Ilustración 11:Grafica de estado de la transacción por categoría de riesgo.

9.3 Modelo de Clasificación

En nuestro análisis y modelado de datos, hemos aplicado diversas técnicas de preprocesamiento y clasificación fundamentales en el campo de la ciencia de datos, la correcta preparación de los datos, incluyendo la transformación y codificación de variables, es esencial para desarrollar modelos efectivos. Es importante entender las características de los datos y aplicar métodos apropiados para su preparación, algo que hemos seguido en nuestro enfoque. La codificación one-hot, por ejemplo, es una técnica crucial para manejar variables categóricas en modelos de aprendizaje automático, permitiendo a los algoritmos procesar eficazmente estos datos sin introducir sesgos indebidos. Este enfoque está en línea con las mejores prácticas recomendadas en 'The Elements of Statistical Learning', donde se enfatiza la necesidad de técnicas de preprocesamiento adecuadas para maximizar la precisión y eficacia de los modelos de clasificación. (Hastie, 2009).

Para construir los modelos, con datos desbalanceados y balanceados, se siguió un proceso de selección y preprocesamiento de características para optimizar el

rendimiento y reducir la complejidad. Las variables auth_results_detail y Risk_business_description fueron eliminadas debido a su alta cardinalidad (muchos valores únicos), lo que generaba complejidad en los modelos. Además, estas variables ya estaban representadas por otras manejables como auth_results_type y category_risk, respectivamente. Asimismo, Las variables MCC y location_id también fueron excluidas por razones similares, ya que su gran número de categorías únicas añadía complejidad a la ejecución de los modelos debido al volumen de datos. La eliminación de estas variables ayudo a reducir la dimensionalidad y en consecuencia la complejidad del modelo, mejorando su desempeño.

Se extrajeron componentes de la fecha de la columna de received_date para incluir información de año, mes y día como características individuales, y se eliminó la columna original de fecha recibida después de extraer esta información.

Las variables seleccionadas para los modelos se agruparon en las siguientes:

- Variables categóricas: país (issuing_country), franquicia (entry_type), tipo de transacción (transaction_type), tipo de tarjeta (card_type), categoría de riesgo (category_risk), año (year), mes (month), y día (day).
- Variables numéricas: monto total (amount_total) y cantidad de transacciones (qty).

Para el preprocesamiento, se utilizaron técnicas para escalar las variables numéricas y codificar las variables categóricas. El uso de escalado estándar garantiza que las variables numéricas tengan una escala comparable. La codificación de una sola categoría se configuró para generar matrices dispersas.

Durante el análisis, se intentó implementar el modelo Naive Bayes utilizando GaussianNB. Sin embargo, este modelo no admite matrices dispersas (sparse matrices) que es lo que produce el OneHotEncoder por defecto. Inicialmente, se cambió la configuración de OneHotEncoder para generar matrices densas, pero

debido a limitaciones de memoria, este enfoque no fue viable. Por lo tanto, se decidió no utilizar Naive Bayes en este análisis.

Además, se intentó implementar el modelo K-Nearest Neighbors (KNN) utilizando un pipeline que incluía la normalización de los datos mediante StandardScaler y la aplicación de KNN con n_neighbors=3. A pesar de estos ajustes, el proceso de validación cruzada resultó en tiempos de ejecución excesivamente largos, incluso cuando se redujo el número de pliegues (cv=3). El algoritmo KNN utiliza un método de búsqueda exhaustiva (brute force) para encontrar los vecinos más cercanos, lo cual resulta ineficiente cuando se trata de datos dispersos y grandes volúmenes de datos como los nuestros. Debido a su incapacidad para manejar eficientemente el volumen y la complejidad de nuestros datos en un tiempo razonable, hemos concluido que KNN no es una opción viable para nuestro análisis. Por estas razones, se ha decidido no utilizar KNN y en su lugar consideraremos otros modelos más eficientes.

Otro modelo de clasificación que se intentó implementar es Gradient Boosting un algoritmo potente y flexible, pero computacionalmente intensivo, especialmente con grandes conjuntos de datos. Esto se debe a que construye los árboles de decisión de manera secuencial, optimizando cada uno para corregir los errores del anterior, lo que resulto en tiempos de entrenamiento prolongados. Por estas razones, se decidió optar por XGBoost, una implementación más avanzada y eficiente de Gradient Boosting, que utiliza múltiples núcleos del procesador para acelerar el entrenamiento en conjuntos grandes y mejora el rendimiento al incorporar técnicas que evitan el sobreajuste. Esto lo hizo más adecuado para nuestro análisis, dado el volumen de datos y los requisitos de tiempo de procesamiento.

Dado lo anterior, aplicamos los siguientes modelos de clasificación: regresión logística, Arboles de decisión, Random Forest y XGBoost, para predecir y caracterizar las transacciones aprobadas y declinadas. Los modelos fueron ajustados utilizando un conjunto de datos para entrenamiento y validados con un conjunto de pruebas independiente. Además, se evaluaron en términos de

precisión, recall y F1-score, para realizar un análisis comparativo con el fin de determinar cuál de los modelos proporcionaba los mejores resultados en términos de precisión y capacidad para identificar transacciones aprobadas o declinadas.

9.4 Modelos de clasificación con datos desbalanceados

En nuestro análisis de datos desbalanceados, se probaron y evaluaron los modelos de clasificación con un conjunto de datos con las siguientes dimensiones:

- Conjunto de entrenamiento: 5,193,044 muestras y 251 características.
- Conjunto de prueba: 1,298,261 muestras y 251 características.

Para validar los modelos de clasificación implementados, se realizaron predicciones mediante validación cruzada utilizando el método predict_proba el cual genera una matriz con las probabilidades estimadas de cada clase, adicional, se calcularon la curva ROC para cada modelo.

A continuación, se presentan los detalles y los resultados de la validación cruzada y las métricas de rendimiento:

9.4.1 Regresión Logística

La regresión logística como se indicó previamente es un método estadístico utilizado para modelar la probabilidad de un resultado binario, es decir, un resultado que solo puede tomar dos valores. En este estudio los resultados son "Aprobado" o "Declinado", este tipo de modelo de regresión se emplea cuando la variable dependiente es categórica y cuyo objetivo es predecir la probabilidad de que una observación pertenezca a una de las categorías, en este caso 0,0 y 1,0.

Para la ejecución de este modelo de clasificación, las variables numéricas amount_total, qty, y transaction_type fueron normalizadas utilizando StandardScaler. Posteriormente, se procedió con la validación de supuestos, donde se identificó un alta multicolinealidad, problema que se resolvió aplicando la técnica

de Análisis de componentes principales (PCA), permitiendo reducir la dimensionalidad del conjunto de datos y mejorar ligeramente la precisión del modelo.

Este modelo se configuró con el solver 'liblinear' debido a su eficiencia computacional para problemas de clasificación binaria y a su capacidad para converger más rápidamente en comparación con el solver 'saga', el cual se había configurado inicialmente, pero después de 1 hora y media, no logró completar la validación cruzada. Para evitar el sobreajuste, se aplicó una regularización L2 (penalty='l2') y se ajustó el parámetro C a 0.1 para añadir más regularización y mejorar la convergencia del modelo. Además, se aumentó el número de iteraciones predeterminado de 100 a 1000 para asegurar la convergencia.

El modelo se entrenó en aproximadamente 2.12 minutos, logrando un score de validación cruzada de 0.9301.

Logistic Regression with PCA Classification Report:

	precision	recall	f1-score	support
0	0.93	1.00	0.96	4829783
1	0.88	0.00	0.00	363261
accuracy			0.93	5193044
macro avg	0.91	0.50	0.48	5193044
weighted avg	0.93	0.93	0.90	5193044

Ilustración 12:Resultado Regresión logística datos desbalanceados

Este modelo muestra una alta precisión para la clase mayoritaria (transacciones aprobadas). Sin embargo, falla en capturar correctamente la clase minoritaria, a pesar de tener una precisión del 88% para las transacciones declinadas, el modelo no predice adecuadamente esta clase, lo que resulta en un f1-score de 0% y un recall del 0%.

9.4.2 Árboles de Decisión

El modelo de árboles de decisión se configuró con el criterio 'Gini' para medir la calidad de las divisiones. La profundidad máxima del árbol se limitó a 20 para prevenir el sobreajuste y reducir la complejidad del modelo. Se estableció un mínimo de 50 muestras para dividir un nodo interno y un mínimo de 10 muestras por nodo de hoja, asegurando así que las divisiones fueran significativas y evitando el sobreajuste.

Este modelo se entrenó en 15.95 minutos y obtuvo un score de validación cruzada de 0.9294.

Decision Tree Classification Report:

	precision	recall	f1-score	support
0	0.93	1.00	0.96	4829783
1	0.43	0.03	0.06	363261
accuracy			0.93	5193044
macro avg	0.68	0.51	0.51	5193044
weighted avg	0.90	0.93	0.90	5193044

Ilustración 13:Resultado Arboles de Decisión datos desbalanceados

En este modelo se observa una ligera mejora en la captura de la clase minoritaria (transacciones declinadas) en comparación con la regresión logística, con un aumento en el f1-score de 0% a 6%. Sin embargo, el modelo aún muestra un bajo recall del 3% para la clase minoritaria, lo que indica que sigue teniendo dificultades para identificar correctamente la mayoría de las transacciones declinadas.

9.4.3 Random Forest

Inicialmente, el modelo de Random Forest se configuró con 100 árboles y una profundidad de 20. Sin embargo, debido a los altos tiempos de entrenamiento y problemas de memoria durante la validación cruzada, se ajustaron los hiperparámetros, reduciendo el número de árboles a 20 y la profundidad máxima a 5, logrando un equilibrio entre tiempo de entrenamiento y rendimiento.

Este ajuste resultó en un tiempo de entrenamiento de 2.84 minutos y un score de validación cruzada de 0.9300.

Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.93	1.00	0.96	4829783
1	0.00	0.00	0.00	363261
accuracy			0.93	5193044
macro avg	0.47	0.50	0.48	5193044
weighted avg	0.86	0.93	0.90	5193044

Ilustración 14: Resultado Random Forest datos desbalanceados

Al igual que en el caso de la regresión logística, este modelo no captura correctamente la clase minoritaria (transacciones declinadas), resultando en un recall del 0% y un f1-score de 0%. Esto indica que el modelo no logra identificar ninguna transacción declinada, a pesar de su alta precisión en la clase mayoritaria.

9.4.4 XGBoost

XGBoost se configuró con 100 estimadores, una tasa de aprendizaje de 0.1 y una profundidad máxima de 5. Este modelo es conocido por su eficiencia y alta capacidad predictiva. El tiempo de entrenamiento fue de 0.62 minutos, y el modelo obtuvo un score de validación cruzada de 0.9301.

XGBoost Classification Report:

	precision	recall	f1-score	support
0	0.93	1.00	0.96	4829783
1	0.78	0.00	0.01	363261
accuracy			0.93	5193044
macro avg	0.86	0.50	0.48	5193044
weighted avg	0.92	0.93	0.90	5193044

Ilustración 15: Resultado XGBoost datos desbalanceados

Este modelo presenta una mejor precisión para la clase minoritaria (transacciones declinadas) en comparación con los otros modelos, con una precisión del 78%, sin embargo, sigue teniendo un recall muy bajo del 0%, lo que significa que no logra identificar correctamente las transacciones declinadas.

9.4.5 Curva ROC

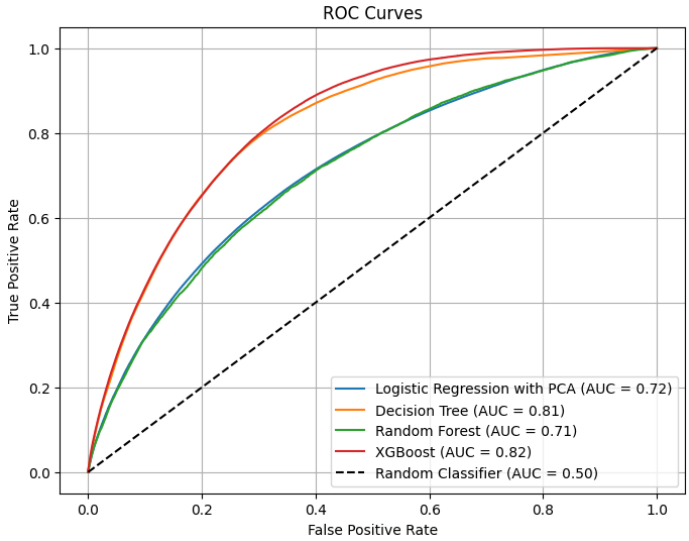

Ilustración 16: Resultado curva ROC datos desbalanceados

En la Ilustración 16, se comparan las curvas ROC (Receiver Operating Characteristic) de los cuatro modelos implementados, cada curva muestra el rendimiento del modelo al variar el umbral de decisión utilizado para clasificar las muestras. En el eje X se presenta la tasa de Falsos Positivos (FPR- transacciones que fueron incorrectamente identificadas como declinadas cuando en realidad deberían haber sido aprobadas) y en el eje Y la tasa de Verdaderos Positivos (TPR- transacciones que fueron correctamente identificadas como declinadas por el modelo).

El área bajo la curva (AUC) de los modelos es la siguiente: Regresión Logística con PCA 0.72, Árboles de Decisión 0.81, Random Forest 0.71 y XGBoost 0.82. Esto indica que entre más se acerca el AUC a 1, mayor es la capacidad del modelo para diferenciar entre las clases positivas y negativas. En este caso, el mejor rendimiento

lo obtuvo el modelo de XGBoost, seguido de cerca por el Árbol de Decisión, lo que sugiere que estos modelos tienen una mejor capacidad para distinguir entre transacciones aprobadas y declinadas.

9.4.6 Análisis de Resultados

Analizando los resultados anteriores, todos los modelos lograron una precisión similar en términos de exactitud, con valores cercanos al 93%. Sin embargo, al observar las métricas de AUC, se evidencian algunas diferencias en la capacidad de los modelos para diferenciar entre las clases, con XGBoost y el Árbol de Decisión mostrando un AUC superior (0.82 y 0.81 respectivamente).

No obstante, las métricas de precisión, recall y F1-score para la clase minoritaria (transacciones declinadas) revelan diferencias significativas. Aunque algunos modelos como XGBoost y la Regresión Logística con PCA muestran una precisión relativamente alta para la clase minoritaria, todos los modelos presentan un recall extremadamente bajo (cercano o igual a 0%), lo que indica una gran dificultad para capturar correctamente las transacciones declinadas.

Estas observaciones indican que, aunque los modelos son efectivos en la clasificación de la clase mayoritaria, su desempeño es insuficiente para la clase minoritaria. Para mejorar la identificación de las transacciones declinadas, es crucial implementar técnicas de balanceo de clases, que permitan aumentar el recall y el F1-score para la clase minoritaria, garantizando así una mejor identificación de transacciones declinadas y una evaluación más equilibrada del modelo.

9.5 Modelos de clasificación con datos balanceados:

Para abordar el problema del desbalance en las clases, implementamos diversas técnicas de balanceo de datos antes de aplicar los modelos de clasificación. La distribución inicial de las clases era la siguiente:

- Clase 0 (Aprobado): 6,036,720 muestras

- Clase 1 (Declinado): 454,585 muestras

Aplicamos las siguientes técnicas de balanceo:

- SMOTE: Genera nuevas muestras sintéticas sin duplicar datos, aunque esta técnica es eficiente para mejorar la calidad del modelo, resultó ser computacionalmente intensiva y no finalizó en un tiempo razonable.
- SMOTEENN: Combina SMOTE para sobre muestreo y ENN para eliminar muestras incorrectamente clasificadas, similar a SMOTE, fue muy costosa en términos de tiempo y memoria.
- Cluster Centroids: Utiliza k-means para agrupar las muestras de la clase mayoritaria y luego reduce el conjunto de datos, esta técnica también resultó ineficiente en términos de tiempo, posiblemente debido a la complejidad del algoritmo k-means.
- Random Undersampling: Reduce el tamaño de la clase mayoritaria seleccionando aleatoriamente una cantidad menor de muestras. Esta técnica demostró ser efectiva y eficiente para nuestro conjunto de datos y fue la que se aplicó para balancear los datos.

Después de balancear los datos, las nuevas dimensiones del conjunto de datos fueron:

- Conjunto de entrenamiento: 727,336 muestras y 251 características
- Conjunto de prueba: 181,834 muestras y 251 características

9.5.1 Regresión Logística

Este modelo se ejecutó de la misma forma que para los datos desbalanceados, realizando normalización a las variables numéricas utilizando StandardScaler y posteriormente realizando la validación de supuestos, donde se identificó un alta multicolinealidad, problema que se resolvió aplicando la técnica de Análisis de componentes principales (PCA), permitiendo reducir la dimensionalidad del conjunto de datos y mejorar ligeramente la precisión del modelo.

De igual forma, se configuró con el solver 'liblinear' debido a su eficiencia computacional y capacidad para converger rápidamente. Se aplicó una regularización L2 y se ajustó el parámetro C a 0.1 para mejorar la convergencia.

El tiempo de entrenamiento fue de 0.36 minutos y el score de validación cruzada fue de 0.6598.

Logistic Regression Classification Report:

	precision	recall	f1-score	support
0	0.66	0.64	0.65	363632
1	0.65	0.67	0.66	363704
accuracy			0.65	727336
macro avg	0.65	0.65	0.65	727336
weighted avg	0.65	0.65	0.65	727336

Ilustración 17:Resultado Regresión logística datos balanceados

El modelo de regresión logística muestra una precisión moderada para ambas clases. Sin embargo, su recall es bastante equilibrado entre las dos clases, con un valor de 0.67 para las transacciones declinadas y 0.64 para las aprobadas. Esto indica que el modelo logra capturar una proporción razonable de las transacciones aprobadas y declinadas, aunque no de manera perfecta.

9.5.2 Árboles de Decisión

Se configuró con el criterio 'Gini', una profundidad máxima de 20, un mínimo de 50 muestras para dividir un nodo interno y un mínimo de 10 muestras por nodo de hoja. El tiempo de entrenamiento fue de 1.75 minutos y el score de validación cruzada fue de 0.7467

Decision Tree Classification Report:

	precision	recall	f1-score	support
0	0.79	0.67	0.73	363632
1	0.71	0.82	0.76	363704
accuracy			0.75	727336
macro avg	0.75	0.75	0.75	727336
weighted avg	0.75	0.75	0.75	727336

Ilustración 18: Resultado Arboles de Decisión datos balanceados

El modelo de árbol de decisión muestra una mayor precisión y recall para las transacciones declinadas (clase 1), con un f1-score de 0.76 para esta clase y 0.73 para las transacciones aprobadas (clase 0). Esto sugiere que el modelo es más efectivo en identificar las transacciones declinadas, con un recall del 82%, lo que indica una mayor capacidad para capturar estas transacciones en comparación con las aprobadas, que tienen un recall del 67%. Aunque el modelo es efectivo en ambas clases, tiene un rendimiento ligeramente superior en la identificación de transacciones declinadas.

9.5.3 Random Forest

Configurado con 50 árboles y una profundidad máxima de 10 para mejorar la eficiencia. El tiempo de entrenamiento fue de 0.55 minutos y el score de validación cruzada fue de 0.6898.

Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.69	0.68	0.69	363632
1	0.69	0.70	0.69	363704
...				
accuracy			0.78	727336
macro avg	0.79	0.78	0.78	727336
weighted avg	0.79	0.78	0.78	727336

Ilustración 19:Resultado Random Forest datos balanceados

El modelo de Random Forest muestra una precisión y un recall más balanceados en comparación con la regresión logística, pero algo menores que los obtenidos con el modelo de árbol de decisión. Con una precisión y un f1-score de 0.69 para ambas clases, este modelo proporciona un rendimiento razonable, equilibrando de manera efectiva la captura de transacciones tanto aprobadas como declinadas. Este balance indica que el modelo tiene un desempeño consistente, aunque no sobresaliente, en la clasificación de ambas clases.

9.5.4 XGBoost

Se configuró con 50 estimadores, una tasa de aprendizaje de 0.1 y una profundidad máxima de 10. El tiempo de entrenamiento fue de 0.15 minutos y el score de validación cruzada fue de 0.7816.

XGBoost Classification Report:

	precision	recall	f1-score	support
0	0.83	0.70	0.76	363632
1	0.74	0.86	0.80	363704
accuracy			0.78	727336
macro avg	0.79	0.78	0.78	727336
weighted avg	0.79	0.78	0.78	727336

Ilustración 20: Resultados XGBoost datos balanceados

XGBoost demuestra ser el modelo más efectivo, con una mayor precisión, recall y f1-score en ambas clases en comparación con los otros modelos. Con una precisión de 0.74 para la clase declinada y 0.83 para la clase aprobada, este modelo proporciona un rendimiento sobresaliente, especialmente en términos de balance entre las dos clases. El recall de 0.86 para las transacciones declinadas destaca la capacidad de XGBoost para capturar correctamente una mayor proporción de esta clase, lo que se traduce en un f1-score de 0.80, superior al de los otros modelos evaluados.

9.5.5 Curva ROC

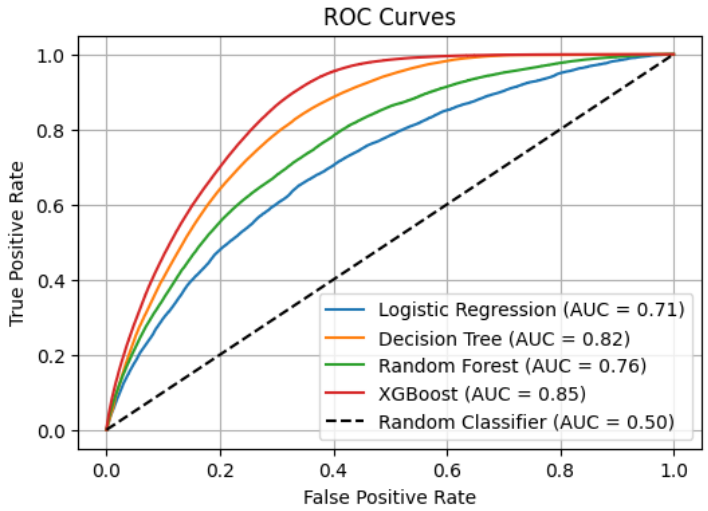

Ilustración 21:Resultado curva ROC datos balanceados

En la Ilustración 21, se comparan las curvas ROC (Receiver Operating Characteristic) de los cuatro modelos implementados. Cada curva muestra el rendimiento del modelo al variar el umbral de decisión utilizado para clasificar las muestras. En el eje X se presenta la tasa de Falsos Positivos (FPR- transacciones que fueron incorrectamente identificadas como declinadas cuando en realidad deberían haber sido aprobadas) y en el eje Y la tasa de Verdaderos Positivos (TPR- transacciones que fueron correctamente identificadas como declinadas por el modelo).

El área bajo la curva (AUC) de los modelos es la siguiente: Regresión Logística 0.71, Árboles de Decisión 0.82, Random Forest 0.76 y XGBoost 0.85. El AUC indica que cuanto más se acerca el AUC a 1, mayor es la capacidad del modelo para diferenciar entre las clases positivas y negativas. Esto sugiere que el mejor rendimiento lo obtuvo nuevamente el modelo de XGBoost.

9.5.6 Análisis de Resultados

Analizando los resultados, todos los modelos lograron mejoras en términos de precisión, recall y F1-score para la clase minoritaria (transacciones declinadas) en

comparación con los datos desbalanceados. Esto resalta la importancia de aplicar técnicas de balanceo de clases, ya que estas técnicas permiten mejorar significativamente la identificación de transacciones declinadas y asegurar una evaluación más equilibrada y justa de los modelos. A continuación, se presentan los datos obtenidos en los modelos de clasificación balanceados:

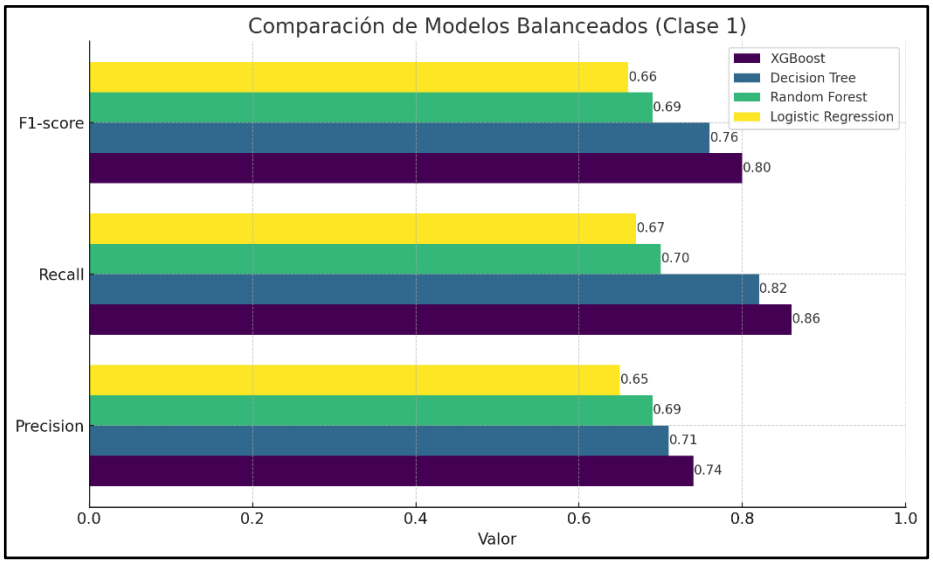

Ilustración 22:Comparación Modelos balanceados (Clase 1-Declinadas)

El preprocesamiento mediante Random Undersampling mejora significativamente el rendimiento de los modelos, equilibrando mejor la precisión y el recall para ambas clases en comparación con los modelos entrenados con datos desbalanceados. Este enfoque demuestra ser eficaz para manejar el desbalance en el conjunto de datos, proporcionando una base sólida para la clasificación de transacciones aprobadas y declinadas. En particular, el modelo de clasificación XGBoost con datos balanceados muestra el mejor rendimiento general, con mejoras notables en la precisión y el F1-score de la clase minoritaria (transacciones declinadas).

9.6 Análisis de Estadística Multivariada

Tal y como se indicaba previamente el algoritmo Isolation Forest es un modelo de aprendizaje automático utilizado para la detección de anomalías, el cual

analizaremos en este proyecto, entrenando el modelo de Isolation Forest en el conjunto de datos procesado, con el objetivo de identificar transacciones atípicas, pues este modelo es capaz de aislar observaciones que se desvían de la norma, lo que lo facilita la identificación de anomalías en los datos.

Para la implementación de este modelo, fue necesario tomar una muestra del 10% del conjunto de datos completo debido que computacionalmente no se logró ejecutar para toda la base (Ejecución 45 minutos). Esta muestra se validó para asegurar su representatividad en comparación con el conjunto de datos original.

Para validar la representatividad de la muestra, se compararon las estadísticas descriptivas de la base inicial con la muestra. Las variables analizadas fueron location_id, mcc, transaction_type, qty, transaction_status y risk. Los resultados indican que las estadísticas descriptivas de estas variables son bastante parecidas en ambos conjuntos de datos, lo que sugiere que la muestra es representativa.

Sin embargo, en la variable amount_total, aunque las medias son bastante cercanas (838.72 en la base completa frente a 825.00 en la muestra), la desviación estándar es considerablemente menor en la muestra (6525.48 frente a 11451.32). Esto sugiere que, aunque hay una diferencia en los valores extremos, la muestra parece capturar adecuadamente la variabilidad de esta variable.

Para confirmar estadísticamente la representatividad de la muestra, se realizó una prueba de hipótesis (prueba t) para comparar las medias entre la base completa y la muestra. Los resultados de esta prueba statistic: 0.950 y p-valor: 0.342 indicaron que no hay una diferencia estadísticamente significativa entre las medias, respaldando así la representatividad de la muestra.

Una vez validada la muestra, se eliminaron las variables Risk_business_description, Category_risk, auth_results_detail y auth_results_type, ya que se encontraban agrupadas en variables incluidas en el análisis. Además, se codificaron las variables entry_type, issuing_country y card_type utilizando get_dummies para convertirlas en variables dummy. Se eliminaron los valores nulos y se procedió a entrenar el

modelo de Isolation Forest, con el objetivo de identificar transacciones que se comportan de manera anómala en comparación con el resto del conjunto de datos. La identificación de estas anomalías es crucial para mejorar la detección de fraudes y errores, así como para optimizar las políticas de auditoría y seguridad en transacciones en línea.

En la aplicación del modelo se estableció un nivel de contaminación del 0.01, lo que implica que se espera que aproximadamente el 1% de las observaciones en el conjunto de datos sean anómalas. El valor de max_samples se fijó en 51,930 observaciones. Como resultado, se identificaron un total de 1,292 alertas tempranas, correspondientes a transacciones que se comportan de manera anómala. Estas alertas son cruciales para la detección temprana de posibles fraudes o errores en las transacciones. De las alertas identificadas, 1,170 corresponden a transacciones aprobadas las cuales suman un total de 4.57 millones de dólares, mientras que 122 corresponden a transacciones declinadas con un monto total de 92 mil dólares.

Estas transacciones pueden ser indicativas de comportamientos inusuales, como posibles inconsistencias o errores, y podrían requerir análisis adicionales para comprender mejor su naturaleza y tomar las medidas adecuadas. La detección de anomalías a través de modelos como Isolation Forest ayuda a las organizaciones a identificar proactivamente posibles problemas o irregularidades en grandes conjuntos de datos, lo cual es esencial para la identificación rápida de transacciones sospechosas.

Aunque modelos como LOF y One-Class SVM son ampliamente utilizados en problemas de detección de anomalías, se optó por no incluirlos en este estudio por varias razones. En primer lugar, durante la prueba se observó que tanto LOF como One-Class SVM presentaron tiempos de entrenamiento significativamente más largos en comparación con el Isolation Forest, lo cual es un factor crucial dado que la eficiencia computacional es fundamental para nuestra investigación. Además, la sensibilidad de estos modelos a la configuración de hiperparámetros y a la

presencia de datos ruidosos puede afectar negativamente su rendimiento en determinadas circunstancias (Liu, Ting, & Zhou, 2008).

9.7 Visualizador en Power BI

El siguiente es el esquema de la arquitectura de datos conectados con Power BI, el cual comienza con "Hundreds of data sources", indicando como Power BI puede conectarse a diferentes fuentes de datos, como BD´s en la nube (Cloud SQL) y aplicaciones como Google Analytics y Salesforce.

Las flechas amarillas que conecta los gateways y el servicio en la nube simboliza el flujo continuo y seguro de datos, manteniendo la información accesible y actualizada para la toma de decisiones basada en datos.

Para asegurar que los datos estén actualizados y sean seguros, se puede utilizar los componentes de "Power BI gateways", los cuales actúan como puentes entre los sistemas de datos locales de la organización y el servicio Power BI en la nube, permitiendo una actualización segura y regular de los datos.

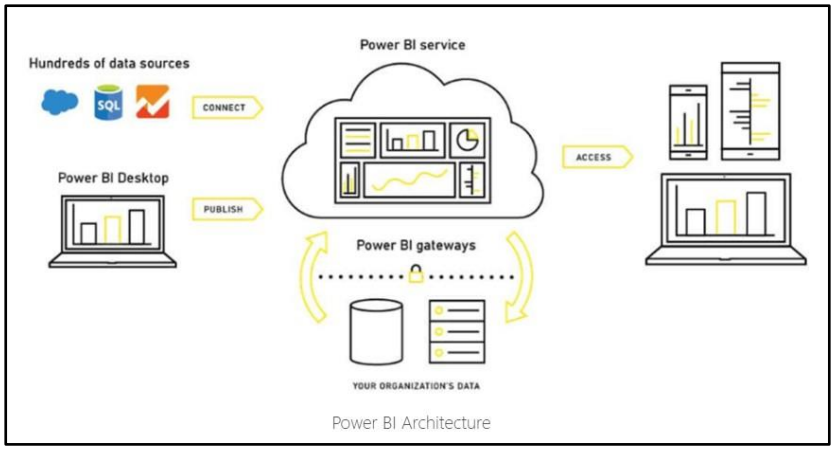

Ilustración 23:Esquema de la arquitectura de datos conectados con Power BI – Fuente Empresa en Estudio

A continuación, se propone el siguiente visualizador, cuyo objetivo es mostrar las variables más influyentes en el proceso de pagos transaccionales y que así los auditores puedan monitorearlo, dando sugerencias oportunas.

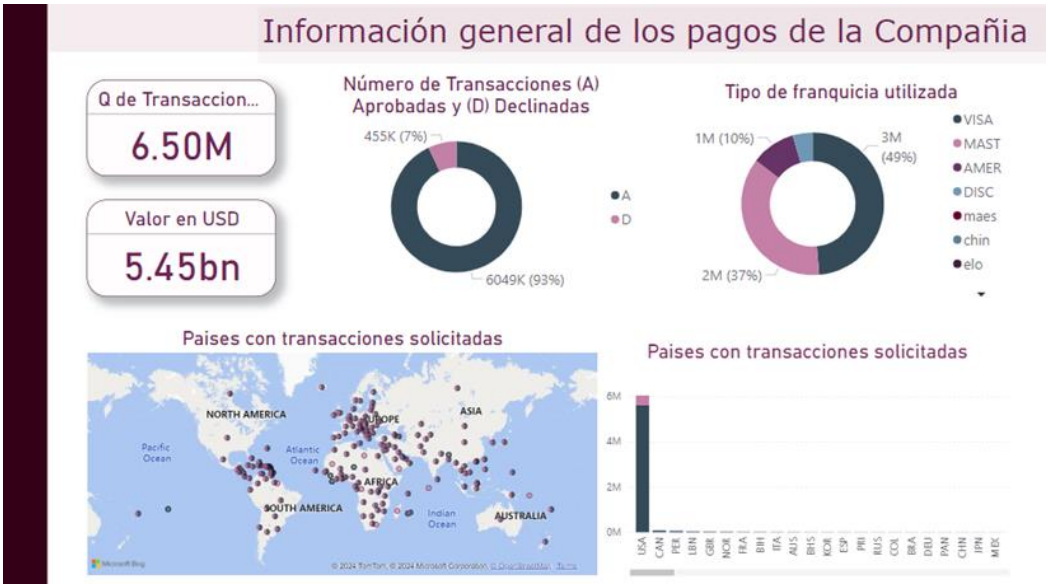

Ilustración 24: Propuesta de Visualizador en Power BI

10. CONCLUSIONES

En respuesta a la pregunta de investigación planteada: **¿Cómo puede la implementación de elementos de analítica y la metodología de auditoría basada en riesgos contribuir a mejorar la eficiencia del proceso de auditoría de pagos transaccionales en línea?,** se identificaron varios aspectos clave que destacan la efectividad de estas herramientas en la optimización del proceso.

Limitaciones de la Metodología de Auditoría Tradicional:

La metodología de auditoría tradicional se basa en analizar una muestra aleatoria de un máximo de 60 transacciones del año anterior, lo cual puede sesgar los resultados y limitar la capacidad del auditor para detectar problemas significativos.

La implementación de un Análisis Exploratorio de Datos (EDA) sobre la totalidad de las bases de datos permitió superar esta limitación, revelando tendencias y factores clave que influyen en las transacciones. Este enfoque proporcionó una comprensión profunda del comportamiento transaccional y facilitó una gestión de riesgos más informada y proactiva.

Hallazgos Críticos y Recomendaciones en el Proceso de Auditoría:

Transacciones en Comercios Prohibidos y Restringidos: Se identificaron 7,587 transacciones aprobadas en comercios clasificados como prohibidos y 4,885 en comercios restringidos. Este hallazgo representa un incumplimiento significativo de los acuerdos con las diferentes franquicias y un incumplimiento de las políticas internas de la compañía. La identificación de estos comercios es crucial para tomar medidas correctivas y asegurar el cumplimiento normativo.

Recomendaciones:

- Implementar controles automáticos en el sistema transaccional que bloqueen de manera preventiva las transacciones a comercios de prohibidos o restringidos, según las políticas internas y los acuerdos con las franquicias.
- Establecer un monitoreo continuo y auditorías periódicas para detectar y detener cualquier transacción en comercios prohibidos o restringidos, informando inmediatamente a los responsables para la toma de decisiones

Transacciones sin Clasificación de Riesgo: Se detectaron aproximadamente 12,000 transacciones sin una clasificación de riesgo asignada debido a la falta de actualización de los códigos MCC (Merchant Category Codes). Esta ausencia de clasificación representa un riesgo importante, toda vez que estas transacciones podrían haber sido declinadas o, al menos, revisadas con mayor detalle. La falta de categorización adecuada podría resultar en disputas, multas o sanciones por parte de las franquicias, por lo que es esencial actualizar y mantener al día los códigos MCC para minimizar riesgos financieros y legales.

Recomendaciones:

- Actualizar inmediatamente la base de datos de los códigos MCC (Merchant Category Codes) y establecer un protocolo de mantenimiento periódico para asegurar que las clasificaciones de riesgo estén siempre actualizadas.
- Implementar un proceso de validación automática que identifique cualquier transacción sin clasificación de riesgo y que detenga su aprobación hasta que la clasificación esté asignada correctamente.

Inconsistencias en el Registro de Transacciones: Durante el análisis, y con el apoyo de los líderes de procesos de la compañía, se identificaron errores significativos en el registro de transacciones en dos fechas clave: el 20 de enero de 2022 y el 29 de agosto de 2022, con montos de 6,786 millones y 1,100 millones de dólares, respectivamente. En ambos casos, varias transacciones fueron agrupadas incorrectamente en una sola, resultando en un rechazo masivo. Este hallazgo subraya la necesidad de mejorar los controles internos y procesos para evitar errores similares en el futuro y garantizar la precisión en el registro de transacciones.

Recomendaciones:

- Establecer un proceso de conciliación diario o semanal que permita la verificación automática de los registros de transacciones en tiempo real, asegurando que las transacciones estén correctamente documentadas antes de su cierre final.
- Implementar un sistema de alertas que notifique automáticamente al equipo de finanzas y de pagos cuando se identifiquen transacciones inusualmente altas o agrupaciones incorrectas, para que sean revisadas antes de su procesamiento.

Evaluación de Modelos de Clasificación

En cumplimiento de los objetivos propuestos, se exploraron y aplicaron modelos de clasificación utilizando bases de datos tanto desbalanceadas como balanceadas. Durante este proceso, se evidenció que el modelo de ensamblado XGBoost, mostro un desempeño superior, destacándose por su alta capacidad predictiva para

diferenciar entre transacciones aprobadas y declinadas. Este modelo fue particularmente eficaz en la identificación de transacciones declinadas que no deberían haber sido rechazadas, lo cual es crucial para la optimización del proceso de auditoría. Sin embargo, uno de los mayores desafíos enfrentados fue el manejo del volumen de datos (aproximadamente 6 millones de registros), debido a la falta de experiencia y recursos computacionales suficientes para procesar los modelos. Para mitigar estos obstáculos, fue necesario implementar técnicas avanzadas de preprocesamiento y optimización de recursos, asegurando que el modelo pudiera entrenarse y evaluarse de manera eficiente a pesar de las limitaciones. Estos esfuerzos no solo mejoraron la precisión del modelo, sino que también sentaron las bases para futuras aplicaciones de analítica avanzada en la auditoría de pagos transaccionales.

Detección de Anomalías con Isolation Forest

La implementación del modelo Isolation Forest fue fundamental para identificar transacciones atípicas de manera eficaz. Este hallazgo permite la detección temprana de fraudes y errores, fortaleciendo la infraestructura de seguridad y mejorando el proceso de auditoría de pagos transaccionales en línea

Desarrollo de un Visualizador en Power BI:

El desarrollo de un visualizador en Power BI representó un avance significativo en la auditoría de pagos. Esta herramienta permitió visualizar los datos de toda la población de transacciones, facilitando la identificación de factores críticos y mejorando la toma de decisiones basada en datos. Además, el visualizador ayuda a resaltar deficiencias de control y a generar alertas tempranas, lo que mejora la capacidad de la empresa para gestionar riesgos de manera oportuna y eficaz.

11. REFERENCIAS BIBLIOGRAFICAS

1. Bowlin, K. (2011). Risk-based auditing, strategic prompts, and auditor sensitivity to the strategic risk of fraud. Accounting Review, 86(4), 1231–1253. https://publications.aaahq.org/accounting-review/article-abstract/86/4/1231/3292/Risk-Based-Auditing-Strategic-Prompts-and-Auditor?redirectedFrom=fulltext
2. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://link.springer.com/article/10.1023/A:1010933404324
3. Calderon, T. G., & Cheh, J. J. (2002). A roadmap for future neural networks research in auditing and risk assessment. International Journal of Accounting Information Systems, 3(4), 203–236. https://www.sciencedirect.com/science/article/abs/pii/S1467089502000684?via%3Dihub
4. Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://www.jair.org/index.php/jair/article/view/10302
5. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://dl.acm.org/doi/10.1145/2939672.2939785
6. Capgemini (2023). World Payments Report 2023. Winning with SMBs: Optimizing technology and data to drive deep engagement. Retrieved from https://www.capgemini.com/insights/research-library/world-payments-report/
7. Deloitte. (2020). "Internal Audit in the Payments Industry". Recuperado de https://www2.deloitte.com/us/en/pages/advisory/articles/internal-audit-payments-industry.html
8. Ding, Z., & Fei, M. (2013). An Anomaly Detection Approach Based on Isolation Forest Algorithm for Streaming Data using Sliding Window. IFAC

Proceedings Volumes, 46(20), 12–17. https://www.sciencedirect.com/science/article/pii/S1474667016314999?via%3Dihub

9. Investopedia. (s.f.). *Accounting Period: What It Is, How It Works, Types, and Requirements*. https://www.investopedia.com/terms/a/accounting-period.asp
10. Corporate Finance Institute. (s.f.). *Reporting Period - Definition, Cycles, Importance, Example*. https://corporatefinanceinstitute.com/resources/knowledge/accounting/reporting-period/
11. Trade That Swing. (s.f.). *Best and Worst Months for the Stock Market - Seasonal Patterns*. https://tradethatswing.com/stock-market-seasonal-patterns/
12. Erickson, B. J., & Kitamura, F. (2021). Magician's corner: 9. performance metrics for machine learning models. *Radiology: Artificial Intelligence*, *3*(3). https://pubs.rsna.org/doi/10.1148/ryai.2021200126
13. Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. In Data Mining: Concepts and Techniques.
14. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Editorial.
15. Kiesow, A., Zarvic, N., & Thomas, O. (2014). Continuous auditing in big data computing environments: Towards an integrated audit approach by using CAATTs. Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft Fur Informatik (GI), P-232.
16. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. 2008 Eighth IEEE International Conference on Data Mining,
17. Liu, Q. (2014). The Application of Exploratory Data Analysis in Auditing. In Rutger Repositories.
18. McKinsey & Company (2021). Global Payments 2021: Transformation amid turbulent undercurrents. Philip Bruno, Olivier Denecker, and Marc Niederkorn. Retrieved from https://www.mckinsey.com/industries/financial-

services/our-insights/global-payments-2021-transformation-amid-turbulent-undercurrents

19. Mitnick, K. (2016). The art of invisibility: The world's most wanted hacker breaks down the hidden techniques to protect your privacy, security, and wealth. New York, NY: Crown Publishing Group. Párrafo 1, página 232.
20. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://www.scopus.com/inward/record.uri?eid=2-s2.0-80555140075&partnerID=40&md5=63e53cee7a9711760872d4d103e5453a
21. Pickett, K. H. S. (2015). Audit Planning: A Risk-Based Approach. In Audit Planning: A Risk-Based Approach.
22. Provost & Fawcett. (2013). Data Science for Business What You Need to Know About Data Mining and Data-Analytic Thinking. In Journal of Chemical Information and Modeling.
23. Rambure, D., & Nacamuli, A. (2008). Payment Systems: From the Salt Mines to the Board Room. Palgrave Macmillan.
24. Zaki, M. J., & Meira Jr., W. (2020). Data mining and machine learning: Fundamental concepts and algorithms (2nd ed.). Cambridge University Press.
25. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, 73, 220-239.
26. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284. https://ieeexplore.ieee.org/document/5128907

27. Loh, W. Y. (2014). Fifty years of classification and regression trees. International Statistical Review, 82(3), 329-348. https://onlinelibrary.wiley.com/doi/10.1111/insr.12016
28. Ding, Z., & Fei, M. (2013). An Anomaly Detection Approach Based on Isolation Forest Algorithm for Streaming Data using Sliding Window. IFAC Proceedings Volumes, 46(20), 12–17.
29. Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. In Data Mining: Concepts and Techniques.